Text Classification for Intelligent Portfolio Management

Young-Woo Seo Joseph Giampapa Katia Sycara

CMU-RI-TR-02-14

May 2002

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2002		**00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Text Classification for Intelligent Portfolio Management	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Robotics Institute,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In the application domain of stock portfolio management, software agents that evaluate the risks associated with the individual companies of a portfolio should be able to read electronic news articles that are written to give investors an indication of the financial outlook of a company. There is a positive correlation between news reports on a company?s financial outlook and the company?s attractiveness as an investment. However, because of the volume of such reports, it is impossible for financial analysts or investors to track and read each one. Therefore, it would be very helpful to have a system that automatically classifies news reports that reflect positively or negatively on a company?s financial outlook. To accomplish this task, we treat the analysis of news articles as a text classification problem. We developed a text classification algorithm that classifies financial news article by using a combination of a reduced but highly informative word feature sets and a variant of weighted majority algorithm. By clustering words represented in latent semantic vector space by LSA into groups with similar concepts, we are able to find semantically coherent word groups. A learning method with unlabeled data ?Self-Confident? sampling was proposed to handle the problem of expensive data labeling. Vote entropy is the criterion that information-theoretically assigns a label to an unlabeled document. In comparison with naive Bayes classification boosted by Expectation Maximization (EM), the proposed method showed a better performance in terms of accuracy. Two criteria are used to evaluate methods: how well they improve their performances with unlabeled data after being initially trained on a small number of human-labeled articles and how well they classify the latest financial news articles which are mostly not seen during the training. The contribution of this work lies in the new classification method that we propose and in the sampling technique we used for improving classification accuracy.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**20**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Abstract

In the application domain of stock portfolio management, software agents that evaluate the risks associated with the individual companies of a portfolio should be able to read electronic news articles that are written to give investors an indication of the financial outlook of a company. There is a positive correlation between news reports on a company's financial outlook and the company's attractiveness as an investment. However, because of the volume of such reports, it is impossible for financial analysts or investors to track and read each one. Therefore, it would be very helpful to have a system that automatically classifies news reports that reflect positively or negatively on a company's financial outlook. To accomplish this task, we treat the analysis of news articles as a text classification problem. We developed a text classification algorithm that classifies financial news article by using a combination of a reduced but highly informative word feature sets and a variant of weighted majority algorithm. By clustering words represented in latent semantic vector space by LSA into groups with similar concepts, we are able to find semantically coherent word groups. A learning method with unlabeled data "Self-Confident" sampling was proposed to handle the problem of expensive data labeling. Vote entropy is the criterion that information-theoretically assigns a label to an unlabeled document. In comparison with naive Bayes classification boosted by Expectation Maximization (EM), the proposed method showed a better performance in terms of accuracy. Two criteria are used to evaluate methods: how well they improve their performances with unlabeled data after being initially trained on a small number of human-labeled articles and how well they classify the latest financial news articles which are mostly not seen during the training. The contribution of this work lies in the new classification method that we propose and in the sampling technique we used for improving classification accuracy.

Contents

1 Introduction

In the application domain of stock portfolio management, there is a large volume of information about a company and its financial performance for humans to effectively attend to and manage while making decisions. To address this problem, we proposed a multi-agent system, called *Warren*[1] [4], [14] that helps the user track information on a portfolio of companies of interest by providing an evaluation of the risks associated with the individual company of interest. Warren is composed of different agents that help the user track the stock price, performance history, earnings summaries, and Beta value (risk) associated with the individual holdings in their stock portfolio, and to proactively advise the user whenever the portfolio may be too risky for the user's preferred tolerance to risk. To supplement the data on a company, the user has access to a *Breaking News agent*, which gathered financial news from on-line news providers such as Reuters, CNN Financial Network, Business Wire, Forbes.com and others. In this paper, we describes our endeavor to create an agent that analyzes news articles that were retrieved by the Breaking News agent for their content about a company's financial well-being, for presentation to the user in a meaningful way.

To accomplish this task, we devised a new text classification algorithm that classifies financial news into the predefined five classes: "good", "good, uncertain", "neutral", "bad, uncertain", and "bad" and a sampling algorithm that automatically assigns the label of unlabeled data information-theoretically. Our hypotheses for how this goal can be achieved are as follows: (1) highly informative words (or phrases) features in a particular class can allow a classifier to estimate the class of a news article with high probability; (2) the *domain experts* algorithm based on the voting process among identified features can perform well for this problem; (3) the performance of the method can be improved by using a sampling method which made use of *vote entropy* as the sampling criteria. Briefly, the proposed method predicts the label of a news articles through voting process among identified domain experts. The domain experts are defined as a set of words describes a particular class relatively well and accordingly help a simple classification algorithm discern the boundary of a class (e.g. "good") from that of another. We made use of two types of domain experts: a set of words highly frequently co-occurred in a particular class and a number of word clusters semantically coherent.

Our task is a text classification problem in that it is to assign an appropriate class label to each given document according to the semantic content of the document. Numerous statistical and machine learning methods have been applied to this domain in recent years including nearest neighbor classification [15], naive Bayes with EM (Expectation Maximization) [11] [13], Winnow with active learning [10], Support Vector Machines (SVMs) [6], [8], Maximum-Entropy model [12]. It is, however, slightly different with others in that our task deals with relatively objective and confined classes such as "good" or "bad" for a company's financial outlook than classification of news articles into "politics" or "economics."

The text classification has several characteristics that make it a difficult domain

[1]The system is named after Warren Buffet, a famous American investor and author about investment strategies.

for the use of machine learning, including a very large number of input features, class noise, and a large percentage of features that are irrelevant. For instance, the exploitation of supervised learning requires a relatively large number of labeled examples. When it is given a small set of labeled training data, classification accuracy will suffer because the variance of data (e.g. difference of vocabulary between training and testing data) will be high. However it is expensive to obtain labeled training data, while unlabeled data are cheaply available. Several methods have been used for coping with the problem of insufficient labeled data, such as Expectation Maximization (EM) [11], [13], selective-sampling [2], sub-sampling and uncertainty sampling [9]. The proposed sampling method, self-confident sampling, picks out least uncertain data from unlabeled data sets in terms of entropy. It is similar to uncertainty-sampling in that it predicts the label of unlabeled data on the basis of the learner's confidence which is acquired during training phase. The examples that are predicted with the least uncertainty will be added to the training set in the next training iteration. The overall procedure of self-confident sampling is described in Section 4.2.

The paper is organized as follows. Section 2 will give the overview of our task in terms of the text classification context. Section 3 details the method of identifying a set of domain-experts for each class. In Section 4, we describes the procedure of classification with consideration of the company's financial well-being. Section 5 provides the experimental results and compares them with those of existing methods. Section 6 discuss the results and the future work respectively.

2 Classification of Financial News Articles

Concisely, our task is to develop an algorithm that classifies each given news article into the predefined classes in terms of the referred company's financial well-being.

The financial news articles gathered for experiments were manually labeled into 5 classes by considering how explicitly they mentioned the company's financial status. The following five classes are considered to be pertinent by considering the nature of financial news articles:

GOOD News scripts which show good evidences of the company's financial status explicitly.
e.g.) ... Shares of ABC Company rose $1/2$ or 2 percent on the Nasdaq to $24-15/16. ...

GOOD, UNCERTAIN News scripts which refer to predictions of future profitability, and forecasts.
e.g.) ... ABC Company predicts fourth-quarter earnings will be high. ...

NEUTRAL News scripts which did not mention anything about the financial well-being of the company explicitly.
e.g.) ... ABC and XYZ Inc. announced plans to develop an industry initiative. ...

BAD, UNCERTAIN News scripts which refer to predictions of future losses, or no profitability.

[id] 000x-xx [\id]
[title] Goldman Profits Fall 13 Percent [\title]
[date] Mar 20 6:35 PM ET [\date]
[source] Reuters [\source]
[company] Goldman Sachs (GS) [\company]
[body]
Goldman Sachs Group Inc.(NYSE:GS - news), one of Wall Street's top firms,
on Tuesday said first-quarter profits fell 13 percent but were above reduced estimates
as fees for advising companies on stock sales declined in a slumping market.
...
The value of Goldman's principal investments fell $140 million, compared with a gain of
$214 million in last year's first quarter. Principal investments were down across the board,
Viniar said.
[\body]
[label] bad [\label]

Figure 1: A example of financial news article.

e.g.) ... ABC (Nasdaq: ABC) warned on Tuesday that Fourth-quarter results could fall short of expectations. ...

BAD News scripts which show bad evidences of the company's financial status explicitly.
e.g.) ... Shares of ABC (ABC: down $0.54 to $49.37) fell in early New York trading. ...

Any news articles that do not mention financial facts of a company explicitly were classified into "neutral" class because it is difficult to determine the company's current financial status. In order to avoid inconsistent assignment of class label to a news article, two "uncertain" classes (e.g., "good, uncertain" and "bad, uncertain") are prepared. One may be allowed to decide it as a good (or bad) news for the company, but we could not be sure of it (i.e. uncertain.) The prediction of future earning is one of examples of these classes because it is a predicted statement, not a description of current fact. Figure 1 shows an example of news article used for our experiments.

3 Identification of Domain Experts

A group of domain experts is defined as a set of words (or phrases) describes a particular class relatively well and accordingly help a classification algorithm discern the boundary of a class from that of another. A domain expert is compatible with a term (n-gram word) feature in other text learning task. However we use this terminology because we focused on identifying a good set of word feature, rather than building a good classification algorithm. We made use of two types of domain experts: a set of words highly frequently co-occurred with a particular class and a word cluster with similar

semantic concepts. They are similar to each other in that both of them are discovered by the concept of word co-occurrence.

3.1 Frequently Co-occurred Phrase

A co-occurred phrase is a word pair that frequently occurred in documents belong to the same class. It is syntactically a sequence of nearby, but not necessarily consecutive words. We believed that a set of such co-occurred phrases discriminates well the class of text documents by themselves without help of a complicated classification if it is strongly associated with its class. For example, *Shares* and *rose* could be a good indicator of "good" class that is selected from a sentence "Shares of Company Acme rose 1/2 or 2 percent on the Nasdaq to $24-15/16..."

It, however, is not easy to select such a set of word pairs due to the inherent complexities of text classification (e.g., large available word features and much noise) and selection of criteria on a strong association between a bigram and a class. To deal with this problem, we first made use of a heuristic that considers the characteristic of financial news report. Since most of financial news articles report several company's stories in a news article, they mentioned a company's name (or a company's ticker[2]) explicitly. From this observation, we built an abridged version of each of news articles. That is, if a sentence contains a company's name or ticker, it is added to the abridged version from the original news text. Indeed, an abridged version of article still has noise, but it also has sufficient information that allow us to determine the correlation between bigrams and a class. A number of bigrams are initially compiled after removing stop-words. To determine a strong association between a bigram and a particular class, the information gain measure was employed [16]. Let $\{c_j\}_{j=1}^{m}$ denote the set of classes in the target space. The information gain of kth bigram candidate in jth class, $bigram_{k,j}$ is defined to be:

$$
\begin{aligned}
Gain(\text{bigram}_{k,j}) &= -\sum_{j=1}^{m} P(c_j)\log P(c_j) \\
&+P(\text{bigram}_{k,j})\sum_{j=1}^{m} P(c_i|\text{bigram}_{k,j})\log P(c_j|\text{bigram}_{k,j}) \\
&+P(\bar{\text{bigram}}_{k,j})\sum_{j=1}^{m} P(c_j|\bar{\text{bigram}}_{k,j})\log P(c_j|\bar{\text{bigram}}_{k,j}) \qquad (1)
\end{aligned}
$$

Equation 1 was applied each of bigram candidates which is made by combining each word in condensed version and five consecutive words toward the end of a sentence. One of the five candidates which has the highest value of information gain is selected for a domain expert for a class. Table 1 shows the example of selected bigrams for each class.

[2] A ticker is a symbol that usually is used for representing a company's name briefly in stock trade market.

class	selected bigrams
GOOD	“revenue rose”, “exceeds expectations”, “share rose”, “rose profit”
GOOD, UNCERTAIN	“expect earnings”, “forecasts earnings”, “anticipate earnings”
NEUTRAL	“alliance company”, “alliance corp”, “introduce ”, “announces product”
BAD, UNCERTAIN	“warning profits”, “short expectation”, “warning earnings”
BAD	“share off”, “share down”, “profit decrease”, “fall percent”, “sales decrease”

Table 1: Examples of selected frequently co-occurred bigrams for each class are represented.

3.2 Semantically Coherent Word Groups

Another method for identifying a set of domain experts is that clusters words (i.e., unigrams) into groups with similar concepts. A word cluster play a role of a domain expert in behalf of all words in the cluster. The word similarity is estimated by co-occurrence between two words in question. A preliminary structure for calculating co-occurrence is an inverted index where each row represents an unigram word and each column does a document in a given document collection. An inverted index is usually represented by word-by-document matrix that each of cell is a frequency of a word in a document. Now the similarity between two words is estimated by computing cosine angle between two word vectors. That is, the more documents that two words are co-occurred, the higher similarity value. However by this way we might be fail to identify a group of word semantically similar due to the sparseness of an inverted index matrix. Accordingly it is not good at capturing “semantic” similarity among words. In other words, a two words in high rate of co-occurrence does not necessarily means that they are semantically similar with each other. In order to capture “semantic” coherence between words, Latent Semantic Analysis was employed (LSA) [5]. LSA is a technique that discovers the salient semantic relationships between words by representing the original word-by-document matrix in a low dimensional linear combination of orthogonal (singular) variables. A matrix decomposition (i.e., Singular Value Decomposition (SVD)) plays a pivotal role to generate a large number of orthogonal singular factors from an inverted index matrix. A small number of the most important singular factors are then selected to approximate the covariance of the original inverted matrix. LSA ultimately captures the “semantic” subject of a given document collection by analyzing the patterns of co-occurrence between words. It is quite useful to deal with the problem of identifying the similarity of documents described in the same subject with different vocabulary, by representing the subject of a document rather than specific words.

Instead of using this semantic representation for directly calculating the similarity between two documents, we made use of this representation scheme as a ground for clustering words under the similar semantic subject. LSA applied to the original in-

verted index for each class to derive an semantically-coherent matrix in terms of the subjects documents describe. A word in the identified vocabulary is represented a word vector. A hierarchical agglomerative clustering [7] is employed to group words.

4 Classifying News by Considering Financial Status

As described earlier, a financial news article is comprised of two groups of sentences: an abridge text and other parts not belong to abridge text. Regardless of a particular classification algorithm, an abridge version of news article is represented by the bag of words model – no location information about each word is available and all the words are independent with one another. A multinomial distribution is assumed for naive Bayes classification and a vector space model is for our proposed model. The abridged news article of each news article is represented as a weight vector:

$$\vec{d_i} =< w_1, w_2, ..., w_k, ..., w_{|T|} >, \tag{2}$$

where w_k is the weight of k th word (or phrase) in i th document vector, d_i, which is made up of $|T|$ number of weights.

4.1 Naive Bayes Classifier with EM

The naive Bayes classification is very popular due to its simple implementation and its theoretical soundness. In a Bayesian learning framework, it is assumed that the text data was generated by a parametric model, and the model parameters are estimated by using training data. By applying Baye rule it predicts the class of a testing document with the highest posterior probability that is one of the values from a computation of the conditional probability of c_j given the particular instances of word features $w_1, ..., w_T$. This classification model has a strong independence assumption that all the attributes w_k are conditionally independent given the value of class c_j and its position in the document.

$$\begin{aligned} Pr(c_j|d_i) &= \arg\max_{c_j \in C} Pr(c_j) \prod_k Pr(w_k|c_j) \\ &\approx \frac{Pr(c_j) \prod_{k=1} Pr(w_k|c_j)}{\sum_{j=1}^{|C|} Pr(c_j) \prod_{k=1} Pr(w_k|c_j)} \\ where, & \\ w_k &= \frac{1 + freq_{k,j}}{|J| + |V|} \end{aligned} \tag{3}$$

where $freq_{k,j}$ is the number of times word t_k is occurred, $|J|$ is the total number of unique words in class j, and $|V|$ is the total number of unique words in data set. This weighting method is called "Laplace smoothing" that is intended to avoid the problem of zero probability by assigning a uniform prior (i.e., $\frac{1}{|V|}$).

A known problem of using naive Bayes classification is that the its performance will be decreased by variance from training data. In other words, when it is given

Input:
Training data $D \leftarrow D_l$, where, D_l: labeled news articles

Feature Selection
Choose the co-located phrases among the current labeled news articles.

Do while performance of classifier improve
1. Generate unlabeled data D_u, where, D_u: unlabeled news articles
   Pick news articles randomly from unlabeled data set
2. Use the current classifier and feature sets to estimate the class of each unlabeled article D_u.
   - $\begin{cases} D_u \leftarrow D_u + \lambda d_i & \text{if } d_i > \theta_{ve} \\ \text{skip } d_i \text{ to the next iteration} & \text{otherwise} \end{cases}$
   - $D \leftarrow D_l + D_u$
3. Re-select feature sets by new data set D.

Output:
A classifier that takes an unlabeled articles and predicts a class label.

Figure 2: A pseudo code for Self-confident sampling.

a small set of labeled training data, the accuracy of classification will suffer because variance in the probability distribution of data would be high. And what is worse, it is expensive to acquire a sufficient number of labeled data for training. In [11], they tried to decrease a variance in exploiting unlabeled data by a combination of a variant of Active Learning and Expectation Maximization (EM). In particular, active learning is used to actively select documents for labeling, then EM assigns probabilistically a label of selected unlabeled document. To be more specific, in the Expectation step, the class of a document $Pr(c_j|d_i)$ is probabilistically estimated by using a current estimate of a class $Pr(c_j)$ derived from a set of unlabeled document. For Maximization step calculates a new maximum likelihood estimates for $Pr(c_j)$ using all the labeled data, both original and probabilistically labeled. Consequently this combination allows a naive Bayes model further to improve its classification accuracy.

4.2 Domain Experts with Self-confident Sampling

Figure 3 describes a classification process by a set of identified domain experts in detail. A group of domain experts for a class is comprised of $|K|$ frequently co-occurred word

$C = \{\text{“good”}, \text{“good, uncertain”}, \text{“neutral”}, \text{“bad, uncertain”}, \text{“bad”}\}$
– $fcp_{k,j}$ denotes k th expert in j th domain (or class),
– $w_{k,j}$ denotes the weight associated with $fcp_{k,j}$,
– E_i denotes the most probable class of i th document.

- For each training example $< d_i, c(d_i) >$
  - Initialize e_j to 0
  - For each domain expert $fcp_{k,j}$
    $e_j \leftarrow e_j + w_{k,j}$ if $fcp_{k,j} \in d_i$
  - Predict
    $E_i(d_i) = \arg\max_j \frac{e_j}{\sum_j \sum_k w_{k,j}}$
  - Update weight
    $w_{k,j} \leftarrow \beta w_{k,j}$ if $e_j \neq c(d_i)$ and $fcp_{k,j} \in e_j$

Figure 3: A pseudo code for Domain-Experts algorithm.

pairs or a word cluster. The actual value of $|K|$ is empirically determined. When it make prediction of a news article, it make use of voting among the group of domain experts and then learns the optimal model by altering the weight associated with each domain expert. One attractive property of the proposed algorithm is that it is able to accommodate inconsistent hypothesis as well as consistent hypothesis. In other words, it does not eliminate a hypothesis that is found to be inconsistent with some training documents, but rather reduces its weight with the degree of β. Since we made use of our own text data set, we can tell that there are little word pair which appears only a class. Domain experts algorithm is similar with Sleeping experts algorithm [1], in that they consider each of selected “word pair” as a consistent domain expert (or hypothesis) to the class. On the contrary Sleeping expert did not allow a classifier to have the inconsistent hypotheses.

The self-confident sampling method which we have proposed in figure 2 shares a property of the uncertainty-sampling [9], in that it predicts the label of an unlabeled data on the basis of the learner’s confidence which is obtained through the training phase. The examples that are labeled with the least uncertain will be added to the training set in the next iteration. Unlike uncertainty-sampling, our method rely only on the vote by each of member of domain experts group, which has knowledge induced from the labeled data. We, however, could not rely on its knowledge completely. In this regards, λ is introduced for regulating the degree of reliance on learner’s experience. Empirically, the proposed sampling method shows the best performance at 70 % confidence.

The class of an unlabeled news article is determined by means of vote entropy. Vote entropy is the entropy of the class label distribution resulting from having each group member deterministically “vote” for its winning class [3]. Let $V(j)$ be the number of

Data	$+$	$+/?$	$+/-$	$-/?$	$-$	Total
Labeled	239	70	526	60	344	1239
Unlabeled	–	–	–	–	–	5000
Total	239	70	526	60	344	6239

Table 2: The number of news articles for each class. Relatively smaller data in "uncertain ($+/?$ and $-/?$)" classes could be explained by the objective contents of financial news article in terms of "good" or "bad."

domain experts which are involved in 'voting' for d_i for the class j:

$$VE(d_i) = -\sum_{j}^{|C|} \frac{V(j)}{|K|} \log \frac{V(j)}{|K|}, \quad \text{if } fcp_k \in d_i, \tag{4}$$

where $|K|$ is the number of domain experts which took parts in voting of ith data, d_i which is ith data from the unlabeled data set.

While the vote entropy is 0 if a number of domain experts participating in the vote belong to the same class, the vote entropy is 1 when the vote committee is consist of an equal number of each class. We found that the value of vote entropy for correct assigning a class to an unlabeled data was less than 0.25, whereas the average entropy for incorrect assignment was greater than 0.7.

5 Experimental Results

In this section, we describe the experimental results of the proposed methods, as compared with conventional methods. As mentioned earlier, experiments were performed using the text data which we had made by ourselves. The labeled financial news articles data set amounts to 1,239 financial news articles gathered from several different online news providers: CNN Financial Network, Forbes, Reuters/Reuters Securities, NewsFactors, Motley Fool, CNet, ZDNet, Morningstar.com, Business Week, AP Financial, Business Wire, PR News Wire, and Associated Press. Table 2 describes the distributions of news articles for each class. The phenomenon that "neutral" class has more data than others could be explained by the fact that larger part of them did not mentioned anything about the company's financial well-being explicitly, but deal with general information about the company.

Experiments aimed to verify the proposed methods in terms of two performance criterion: how well it make use of unlabeled data for improving classification accuracy and how accurately it classifies the latest news articles into predefined classes. Firstly, we evaluated whether the proposed sampling method would improve classification performance rates better than those trained by conventional methods. The experiment was performed to show the performance of domain experts with self-confident sampling, naive Bayes with EM, domain experts with EM and naive Bayes with self-confident sampling. Through the experiments, about 25% of the labeled data was used for testing and the rest of labeled set were used to train classifiers. A number of domain

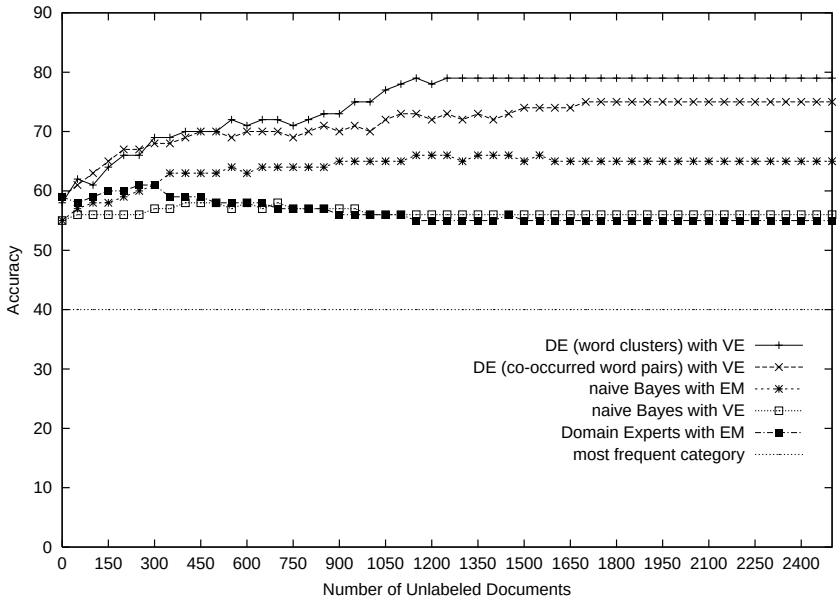

Figure 4: A result of sampling experiment was represented after training each of methods with 690 labeled data. The "most frequent category" is a base line of performance. Since about 40% (526/1239) news articles out of labeled data are "neutral" class, we can assume that a method is able to gain 40% accuracy when it answers consistently the label of each news article in test set as "neutral."

experts for a group is empirically determined by 200 word pairs (i.e., bigrams). After training phase, each methods was tested in terms of classification accuracy: the proportion of the number of news articles classified correctly to the number of total news articles used.

Figure 4 and 5 show results of testing the accuracy performance of each sampling method with different number of labeled data. Total 50 trials were carried out for each method. At each trial, 50 unlabeled news article were given to each methods. When 690 out of 1239 labeled data feeds on training, the performance of the proposed method, the combination of Domain Experts and Self-Confident Sampling, is going up until making use of 1,750 unlabeled news articles, and shows the best performance on accuracy measure at the point. From this, we assumed that around 2,000 news articles allow us to make a classifier with 75 % accuracy because it seems to largely depend on the fact that most of news providers delivered financial news with a restricted vocabulary set. With self-confident sampling, 16% accuracy is improved with 56 % of labeled data (690/1239) and 35% of unlabeled data (1,750/5000) from the result in figure 4. As another goal of our task is to classify the label of the on-line financial news articles, the second experiment was performed to show the accuracy of the latest financial news data. A online data set is made up of the articles that gathered from the same news sources as the labeled data set and reports the latest financial news at the experimental time. At each trial, 30 news articles for a company was gathered from various news sources. However news article about a minor company could not meet the number of

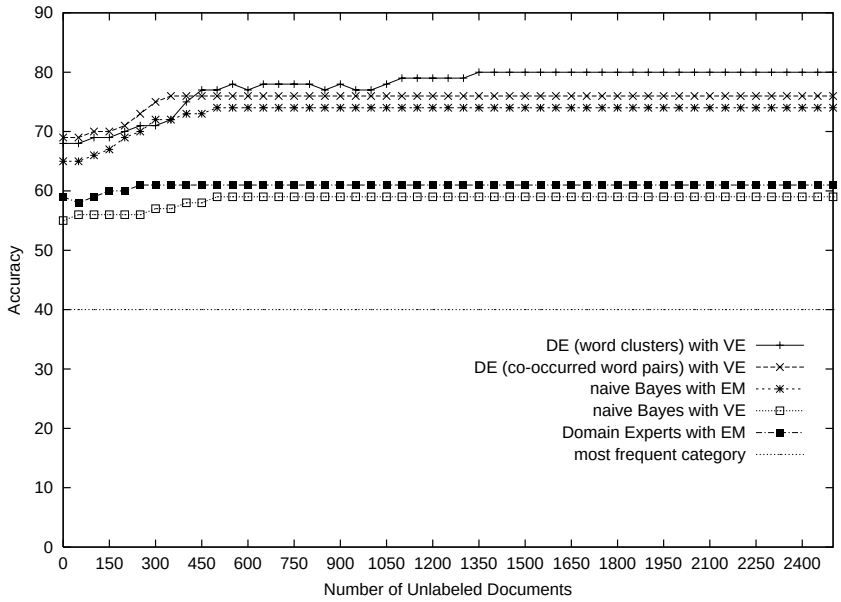

Figure 5: A result of sampling experiment was represented after training each of methods with 1239 labeled data.

Classes	$+$	$+/?$	$+/-$	$-/?$	$-$	Total
Articles	85	1	243	0	220	549
DE (word clusters)	.79	1	.8	–	.85	.82
DE (co-occurred word pairs)	.76	1	.8	–	.78	.79
Naive Bayes	.61	0	.68	–	.62	.65

Table 3: Accuracy measure of each class to the online news data. Each column at the third and fourth row represents the accuracy of each class in terms of the proportion of the number of news article classified correctly to the total number of news article for the class. The column about "good, uncertain ($+/?$)" class has the value of 1 means that only one news article which is labeled by human to that class was classified correctly. There is no news articles about "bad, uncertain ($-/?$)" class.

test documents at a trial (i.e. 30). The second row of Table 3 tells us the distribution of online test set. As a result, the proposed method has 79% averaged accuracy, which means 433 out of 549 total financial news articles were classified correctly. Table 3 shows the accuracy of tested methods per each class.

6 Conclusions

We introduced an application of text classification that classifies financial news articles by considering referred company's financial well-being from their contents. The proposed algorithm which observed co-located phrase of a certain class from news

contents and predicted the label with Weighted-Majority voting outperformed naive Bayes classifier about 14 %. For further improvement of accuracy, we proposed a sampling technique of which determine the class of an unlabeled news article with its entropy value. With the proposed sampling method, self-confident sampling, 16% accuracy is improved with 56 % of labeled data (690/1239) and 35% of unlabeled data (1,750/5000). The successful results from sampling test and online test supports that the proposed algorithms effectively works in this task, even though the promising results are partially come from the task characteristics of which its decision boundaries are relatively objective and are confined with a specific company's name.

But the proposed method has several weak points that prevent it from reaching the performance above 75 % accuracy. One is the difficulty in determining the label of news article of which made up of commensurate number of co-located phrases of each class. To illustrate, "Shares of company B rose 5 % in contrast with company A of which shares fall 7 %." In this example, domain experts may fail to predict "good" for company B. Because both phrases, which are "shares" with "rose" and "shares" with "fall", are very strong indicators of company's financial well-being at the moment, even though they did not indicate the same company and are not assigned with the same weight value during the training phase. As mentioned earlier, another weak point is that the proposed method does not consider the co-referred sentence. In other words, that it does not consider sentences, which did not mention company's name or ticker explicitly, as the financial evidence. For example, "Company C expects to boost revenue next quarter, Chief Operating Officer xxx said Wednesday. Despite of these anticipation, the company's shares fall again." In here, the prediction by the proposed method could be "good, uncertain", even though the true label of this example might be "bad" because "Company C" and "the company" are co-referred as the company's current financial well-being is not good.

To cope with these problems, we consider to employ several natural language processing techniques, such as the consideration of more wide range of a sentence and resolution of co-reference. For the purpose of verifying the applicability of proposed method, we also are about to try to apply the proposed method to the domains of which has similar characteristics to our task.

7 Acknowledgements

We would like to thank Massimo Paolucci and Sean Owens for their maintenance of the Warren financial portfolio agent system. This research has been sponsored in part by DARPA Grant F-30602-98-2-0138 and the Office of Naval Research Grant N-00014-96-16-1-1222.

References

[1] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pages 307–315, 1996.

[2] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[3] I. Dagan and P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of International Conference on Machine Learning*, 1995.

[4] K. Decker, K. Sycara, A. Pannu, and M. Williamson. Designing behaviors for information agents. In *Proceedings of International Conference on Autonomous Agents*, pages 404–413, 1997.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[6] S. Dumais. Using svms for text categorization. *IEEE Intelligent Systems*, 13(4), 1998.

[7] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of European Conference on Machine Learning*, pages 137–142, 1998.

[9] D. Lewis and W. Gale. Training text classifiers by uncertainty sampling. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[10] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[11] A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of International Conference on Machine Learning*, pages 359–367, 1998.

[12] K. Nigam, J. Lafferty, and McCallum. Using maximum entropy for text categorization. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[13] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.

[14] K. Sycara, K. Decker, and A. Pannu. Distributed intelligent agents. *IEEE Expert*, 1996.

[15] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pages 42–49, 1999.

[16] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of International Conference on Machine Learning*, pages 412–420, 1997.